

k -Means is a Variational EM Approximation of Gaussian Mixture Models

Jörg Lücke and Dennis Forster
Machine Learning Group
Carl von Ossietzky Universität Oldenburg
26111 Oldenburg
Germany

April 18, 2017

Abstract

We show that k -means (Lloyd’s algorithm) is equivalent to a variational EM approximation of a Gaussian Mixture Model (GMM) with isotropic Gaussians. The k -means algorithm is obtained if truncated posteriors are used as variational distributions. In contrast to the standard way to relate k -means and GMMs, we show that it is not required to consider the limit case of Gaussians with zero variance. There are a number of consequences following from our observation: (A) k -means can be shown to monotonously increase the free-energy associated with truncated distributions; (B) Using the free-energy, we can derive an explicit and compact formula of a lower GMM likelihood bound which uses the k -means objective as argument; (C) We can generalize k -means using truncated variational EM, and relate such generalizations to other k -means-like algorithms. In general, truncated variational EM provides a natural and quantitative link between k -means-like clustering and GMM clustering algorithms which may be very relevant for future theoretical as well as empirical studies.

1 Introduction

Clustering is the task of associating a set of N data points with a set of C clusters (typically with $C \ll N$), where such an association is defined by a high similarity of points within one cluster compared to the similarity of any two points of different clusters. Different criteria for data point similarity and different algorithmic properties have lead to the development of a large variety of clustering algorithms in the course of more than half a century. Two of the presumably most influential classes of algorithms are

k -means-like algorithms (or Lloyd’s algorithm, Lloyd, 1982) and Gaussian Mixture Models (GMMs).

k-means. k -means or “ k -means like” algorithms (e.g. Steinley, 2006) have been used since the 1950’s and are often considered as the most popular clustering algorithms (Berkhin, 2006). If we denote by $\vec{y}^{(1:N)} = \vec{y}^{(1)}, \dots, \vec{y}^{(N)}$ the data points and by $\vec{\mu}_{1:C} = \vec{\mu}_1, \dots, \vec{\mu}_C$ the cluster centers, a standard form of the k -means algorithm is given by:

Algorithm 1: k -means.

```

initialize  $\vec{\mu}_{1:C}$ ;
repeat
  for  $c = 1, \dots, C$  and  $n = 1, \dots, N$  do
     $s_c^{(n)} = \begin{cases} 1 & \text{if } \forall c' \neq c : \\ & \|\vec{y}^{(n)} - \vec{\mu}_c\| < \|\vec{y}^{(n)} - \vec{\mu}_{c'}\| \\ 0 & \text{otherwise;} \end{cases}$ 
    for  $c = 1, \dots, C$  do
       $\vec{\mu}_c = \frac{\sum_{n=1}^N s_c^{(n)} \vec{y}^{(n)}}{\sum_{n=1}^N s_c^{(n)}}$ ;
until  $\vec{\mu}_{1:C}$  have converged;
```

We will only consider the standard Euclidean metric for $\|\cdot\|$ here. Alg. 1 can be shown (e.g. Inaba and Katoh, 2000) to strictly monotonously increase the k -means objective given by:

$$\mathcal{J}(\vec{\mu}_{1:C}) = \sum_{n=1}^N \sum_{c=1}^C s_c^{(n)} \|\vec{y}^{(n)} - \vec{\mu}_c\|^2. \quad (1)$$

The updates of $s_c^{(n)}$ and $\vec{\mu}_c$ in Alg. 1 are usually derived from (1). Because of its compactness, k -means is easy to implement, and has been observed to work very well in practice (e.g. Duda et al., 2001), especially if improved algorithms, e.g. using D^2 seeding,

are used (e.g. Arthur and Vassilvitskii, 2006; Bachem et al., 2016).

GMM. GMM-based clustering algorithms (e.g. McLachlan and Basford, 1988) are derived from a probabilistic data model, which is in its generative form given by:

$$p(c|\Theta) = \pi_c \quad \text{with} \quad \sum_{c=1}^C \pi_c = 1, \quad (2)$$

$$p(\vec{y}|c, \Theta) = |2\pi\Sigma_c|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|\vec{y} - \vec{\mu}_c\|_{\Sigma_c}^2\right), \quad (3)$$

$$\|\vec{y} - \vec{\mu}_c\|_{\Sigma_c}^2 = (\vec{y} - \vec{\mu}_c)^T \Sigma_c^{-1} (\vec{y} - \vec{\mu}_c) \quad (4)$$

where $\pi_c, \sigma^2 \geq 0$, where Σ_c is for each c a positive definite matrix, and where $|\cdot|$ denotes the determinant. The most standard form to update the model parameters $\Theta = (\pi_{1:C}, \vec{\mu}_{1:C}, \Sigma_{1:C})$ is derived using expectation maximization (EM; Dempster et al., 1977), which results in Alg. 2 (see, e.g., Bishop, 2006).

Algorithm 2: EM for GMM.

init $\Theta = (\pi_{1:C}, \vec{\mu}_{1:C}, \Sigma_{1:C})$;

repeat

for $c = 1, \dots, C$ and $n = 1, \dots, N$ **do**

$$r_c^{(n)} = \frac{\pi_c |2\pi\Sigma_c|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|\vec{y}^{(n)} - \vec{\mu}_c\|_{\Sigma_c}^2\right)}{\sum_{c'} \pi_{c'} |2\pi\Sigma_{c'}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|\vec{y}^{(n)} - \vec{\mu}_{c'}\|_{\Sigma_{c'}}^2\right)}$$

for $c = 1, \dots, C$ **do**

$$\vec{\mu}_c = \frac{\sum_{n=1}^N r_c^{(n)} \vec{y}^{(n)}}{\sum_{n=1}^N r_c^{(n)}}; \quad (\text{GMM 1})$$

$$\Sigma_c = \frac{\sum_{n=1}^N r_c^{(n)} (\vec{y}^{(n)} - \vec{\mu}_c)(\vec{y}^{(n)} - \vec{\mu}_c)^T}{\sum_{n=1}^N r_c^{(n)}}; \quad (\text{GMM 2})$$

$$\pi_c = \frac{1}{N} \sum_{n=1}^N r_c^{(n)}; \quad (\text{GMM 3})$$

until parameters Θ have converged;

The algorithm maximizes (and is derived from) the (logarithmic) data likelihood given by:

$$\mathcal{L}(\Theta) = \frac{1}{N} \sum_{n=1}^N \log \left(\sum_{c=1}^C \pi_c \mathcal{N}(\vec{y}^{(n)}; \vec{\mu}_c, \sigma^2 \mathbb{I}) \right), \quad (5)$$

with $\mathcal{N}(\vec{y}^{(n)}; \vec{\mu}_c, \sigma^2 \mathbb{I})$ as given in (3). Note that (5) is normalized by the number of data points for this study. As is customary for GMMs, we refer to the posteriors $p(c|\vec{y}^{(n)}, \Theta)$ as *responsibilities* and abbreviate them by $r_c^{(n)}$. Computing the $r_c^{(n)}$ in Alg. 2 is referred to as *E-step*, while the parameter updates in Alg. 2 are referred to as *M-step*.

Related Work and Our Contribution. The popularity of k -means and GMM algorithms has resulted

in many theoretical as well as empirical studies of their functional and theoretical properties. More progress using novel versions could be made and much insight could be gained for k -means (Har-Peled and Sadri, 2005; Arthur and Vassilvitskii, 2006; Arthur et al., 2009; Bachem et al., 2016) and GMMs (Chaudhuri et al., 2009; Kalai et al., 2010; Moitra and Valiant, 2010; Belkin and Sinha, 2010; Xu et al., 2016) relatively recently. Because of their similarity, k -means and GMMs have long been formally related to each other. It was thus well-known (see, e.g. MacKay, 2003; Bishop, 2006, and refs. therein) that k -means (Alg. 1) can be obtained as a limit case of EM for GMMs. For this limit a GMM data model with isotropic and equally weighted Gaussians is considered:

$$p(c|\Theta) = \frac{1}{C} \quad (6)$$

$$p(\vec{y}|c, \Theta) = (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{1}{2\sigma^2}\|\vec{y} - \vec{\mu}_c\|^2\right), \quad (7)$$

i.e., setting $\pi_c = \frac{1}{C}$ and $\Sigma = \sigma^2 \mathbb{I}$ in (2) and (3), respectively. If now the limit to zero variances is taken, $\sigma^2 \rightarrow 0$, then the k -means algorithm (Alg. 1) is recovered from the EM algorithm for GMMs (Alg. 2).

The assignment of data points to clusters in k -means is sometimes referred to as ‘hard’, whereas the assignment is ‘soft’ in EM for GMMs. Hard assignments have been considered disadvantageous as the relative importance of the clusters for the data points is not taken into account. Different k -means generalizations have therefore been suggested with the aim to enhance k -means (e.g., faster convergence Har-Peled and Sadri, 2005), to relax its ‘hard’ cluster assignment (e.g. Bezdek, 1981; MacKay, 2003) or to use methods similar to k -means for non-parametric clustering, e.g., by using the k -means/GMM relation $\sigma^2 \rightarrow 0$ (Kulis and Jordan, 2011).

In this work, we show that the k -means algorithm is equivalent to a variational EM algorithm for GMMs with any finite σ^2 . Variational EM seeks to optimize a lower bound (the free-energy) of the data log-likelihood by making use of variational distributions that approximate a-posteriori probabilities. For our study, we use truncated posteriors (Lücke and Sahani, 2008; Lücke and Eggert, 2010) as variational distributions in a fully variational formulation (Lücke, 2016). By regarding k -means as a variational approximation, GMMs and k -means can be quantitatively related without taking a limit to zero cluster variances. Furthermore, the observation that

k -means is a variational optimization implies that it optimizes a lower bound of a GMM log-likelihood. Hence, we can derive free-energy expressions that quantify the link between the k -means and GMM objectives, and as such closely link these two central classes of clustering algorithms. Such a quantitative link has previously (to the knowledge of the authors) not been established nor were any of the corresponding free-energy results derived before.

Truncated approaches have been applied to mixture models before. Work by Dai and Lücke (2014) used truncated approximations for a position invariant mixture model, Forster and Lücke (2017) used truncated approximations for a hierarchical Poisson mixture, and Shelton et al. (2014) and Hughes and Sudderth (2016) used truncated approximations for standard GMMs. However, none of these contributions relate GMMs to k -means or provided any of the other theoretical results derived in this study.

2 Truncated Variational EM and GMMs

The basic idea of truncated EM is the use of truncated approximations to exact posterior distributions (e.g. Lücke and Eggert, 2010; Sheikh et al., 2014). In the notation as used for GMMs above, the truncated approximation takes the form:

$$p(c | \vec{y}^{(n)}, \Theta) = r_c^{(n)} \approx q_c^{(n)} = \frac{p(c, \vec{y}^{(n)} | \Theta)}{\sum_{c' \in \mathcal{K}^{(n)}} p(c', \vec{y}^{(n)} | \Theta)} \delta(c \in \mathcal{K}^{(n)}), \quad (8)$$

where $\mathcal{K}^{(n)}$ is a set of cluster indices (containing different clusters c associated with data point $\vec{y}^{(n)}$). The set of all $\mathcal{K}^{(n)}$ we denote by \mathcal{K} , i.e., $\mathcal{K} = (\mathcal{K}^{(1:N)})$. As is customary for truncated distributions (Lücke and Eggert, 2010; Dai and Lücke, 2014; Shelton et al., 2014; Hughes and Sudderth, 2016), we take the sizes of all $\mathcal{K}^{(n)}$ to be equal, $|\mathcal{K}^{(n)}| = C'$, with $C' \leq C$. The truncated approximation (8) is a good approximation if $\mathcal{K}^{(n)}$ contains all those clusters with significant posterior mass $p(c | \vec{y}^{(n)}, \Theta)$ (i.e., significant non-zero responsibilities $r_c^{(n)}$). Intuitively, truncated approaches can represent very accurate approximations for many data sets, as typically most responsibilities are negligible.

In order to derive a learning algorithm for GMMs based on truncated distributions, we have to answer

the question how the parameters $\mathcal{K}^{(n)}$ and Θ are to be updated. For our purposes we will here make use of a recent study which has addressed this question for general models (with discrete latents) by embedding truncated distributions into a fully variational optimization framework (Lücke, 2016). More specifically, we will use the result of (Lücke, 2016) that a lower bound of the data likelihood (the free-energy) is monotonously increased if: (A) the parameters Θ are updated using standard M-steps with exact posteriors being replaced by truncated posteriors; and (B) that the sets $\mathcal{K}^{(n)}$ can be found using a computationally tractable expression of the free-energy.

For GMMs this means that we can use the standard M-steps of Alg. 2 and replace $r_c^{(n)}$ with the truncated approximations $q_c^{(n)}$ in (8). For the GMM (6) and (7), the truncated responsibilities and M-steps are thus:

$$q_c^{(n)} = \frac{\exp(-\frac{1}{2\sigma^2} \|\vec{y}^{(n)} - \vec{\mu}_c\|^2)}{\sum_{c' \in \mathcal{K}^{(n)}} \exp(-\frac{1}{2\sigma^2} \|\vec{y}^{(n)} - \vec{\mu}_{c'}\|^2)} \delta(c \in \mathcal{K}^{(n)}) \quad (9)$$

$$\vec{\mu}_c^{\text{new}} = \frac{\sum_{n=1}^N q_c^{(n)} \vec{y}_n}{\sum_{n=1}^N q_c^{(n)}} \quad (10)$$

$$\sigma_{\text{new}}^2 = \frac{1}{DN} \sum_{n=1}^N \sum_{c=1}^C q_c^{(n)} \|\vec{y}^{(n)} - \vec{\mu}_c^{\text{new}}\|^2 \quad (11)$$

The parameters $\mathcal{K}^{(n)}$ of the truncated distributions $q_c^{(n)}$ have to be found in the variational E-step. In order to do so, we use the free-energy expression derived in (Lücke, 2016, Prop. 3), which takes for the GMM (6) and (7) the following form:

$$\begin{aligned} \mathcal{F}(\mathcal{K}, \Theta) &= \frac{1}{N} \sum_{n=1}^N \log \left(\sum_{c \in \mathcal{K}^{(n)}} p(c, \vec{y}^{(n)} | \Theta) \right) \\ &= \frac{1}{N} \sum_{n=1}^N \log \left(\sum_{c \in \mathcal{K}^{(n)}} \frac{1}{C} \mathcal{N}(\vec{y}^{(n)}; \vec{\mu}_c, \sigma^2 \mathbb{1}) \right). \end{aligned} \quad (12)$$

The truncated variational E-step (TV-E-step) first optimizes $\mathcal{F}(\mathcal{K}, \Theta)$ w.r.t. \mathcal{K} and the obtained truncated responsibilities $q_c^{(n)}$ are then used in the M-step (10) and (11) to optimize $\mathcal{F}(\mathcal{K}, \Theta)$ w.r.t. Θ . The form of the free-energy (12) and the result that it is monotonously increased by iterating TV-E-step and M-step are the crucial theoretical results in (Lücke, 2016) that are used in this study. Neither of these two results is straight-forward: (A) truncated distributions themselves depend on the model parameters

Θ , and (B) it requires a number of derivation exploiting specific properties of truncated distributions to obtain the compact expression (12).

The TV-E-step now requires finding sets $\mathcal{K}^{(n)}$ which increase $\mathcal{F}(\Theta, \mathcal{K})$. The free-energy (12) is computationally tractable, so a new \mathcal{K} could in principle be found by directly comparing $\mathcal{F}(\mathcal{K}^{\text{new}}, \Theta)$ of a new \mathcal{K}^{new} with $\mathcal{F}(\mathcal{K}^{\text{old}}, \Theta)$ of the current. We can slightly reformulate the problem by considering a specific data point n and cluster $\tilde{c} \in \mathcal{K}^{(n)}$ for which we ask when a new cluster $c \notin \mathcal{K}^{(n)}$ would increase the free-energy $\mathcal{F}(\mathcal{K}^{\text{new}}, \Theta)$. By virtue of the properties of GMM (6) and (7) and due to the specific structure of the free-energy (concavity of the logarithm and the summation of probabilities in Eqn. 12), we can then show:

Proposition 1

Consider the GMM (6) and (7) and the free-energy (12) for $n = 1 : N$ data points $\vec{y}^{(n)} \in \mathbb{R}^D$. Furthermore, consider for a fixed n the replacement of a cluster $\tilde{c} \in \mathcal{K}^{(n)}$ by a cluster $c \notin \mathcal{K}^{(n)}$. Then the free-energy $\mathcal{F}(\mathcal{K}, \Theta)$ increases if and only if

$$\|\vec{y}^{(n)} - \vec{\mu}_c\| < \|\vec{y}^{(n)} - \vec{\mu}_{\tilde{c}}\|. \quad (13)$$

Proof

First observe that the free-energy is increased if $p(c, \vec{y}^{(n)} | \Theta) > p(\tilde{c}, \vec{y}^{(n)} | \Theta)$ because of the summation over c in (12) and because of the concavity of the logarithm. Analogously, the free-energy stays constant or decreases for $p(c, \vec{y}^{(n)} | \Theta) \leq p(\tilde{c}, \vec{y}^{(n)} | \Theta)$. If we use the GMM (6) and (7), we obtain for the joint:

$$p(c, \vec{y} | \Theta) = \frac{1}{C} (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{1}{2\sigma^2} \|\vec{y} - \vec{\mu}_c\|^2\right).$$

The first two factors are independent of the data point and cluster. The criterion for an increase of the free-energy can therefore be reformulated as follows:

$$\begin{aligned} p(c, \vec{y} | \Theta) &> p(\tilde{c}, \vec{y} | \Theta) \\ \Leftrightarrow \exp\left(-\frac{1}{2\sigma^2} \|\vec{y} - \vec{\mu}_c\|^2\right) &> \exp\left(-\frac{1}{2\sigma^2} \|\vec{y} - \vec{\mu}_{\tilde{c}}\|^2\right) \\ \Leftrightarrow -\frac{1}{2\sigma^2} \|\vec{y} - \vec{\mu}_c\|^2 &> -\frac{1}{2\sigma^2} \|\vec{y} - \vec{\mu}_{\tilde{c}}\|^2 \\ \Leftrightarrow \|\vec{y} - \vec{\mu}_c\| &< \|\vec{y} - \vec{\mu}_{\tilde{c}}\|. \end{aligned}$$

□

Proposition 1 means that we have to replace clusters in $\mathcal{K}^{(n)}$ that are relatively distant from $\vec{y}^{(n)}$ by those closer to $\vec{y}^{(n)}$ (and not yet in $\mathcal{K}^{(n)}$). Any such procedure increases the free-energy $\mathcal{F}(\mathcal{K}, \Theta)$ and gives with M-step (10) and (11) rise to a variational

EM algorithm that monotonously increases the lower bound (12) of the data likelihood. Note in this respect that those elements of $\mathcal{K}^{(n)}$ that are not changed, are remembered across TV-EM iterations. The degree how much $\mathcal{F}(\mathcal{K}, \Theta)$ is increased or how long one should seek new clusters in the TV-E-step is a design choice of the algorithm. For a general generative model, $\mathcal{F}(\mathcal{K}, \Theta)$ would be increased partially because full optimization is often computationally infeasible. In the case of GMMs (and other mixture models) we can exhaustively enumerate all clusters, however. For our purposes it is therefore interesting to ask when $\mathcal{F}(\mathcal{K}, \Theta)$ is fully maximized.

Corollary 1

Prerequisites as for Proposition 1. The free-energy $\mathcal{F}(\mathcal{K}, \Theta)$ is maximized w.r.t. \mathcal{K} (with fixed Θ) if and only if for all n the set $\mathcal{K}^{(n)}$ contains the C' clusters closest to data point $\vec{y}^{(n)}$.

Proof

We assume that there are no equal distances among all pairs of data points and cluster centers. If $\mathcal{K}^{(n)}$ contains the C' closest clusters it applies: $\forall c \in \mathcal{K}^{(n)} \forall \tilde{c} \notin \mathcal{K}^{(n)}: \|\vec{y}^{(n)} - \vec{\mu}_c\| < \|\vec{y}^{(n)} - \vec{\mu}_{\tilde{c}}\|$. If we now consider an arbitrary n and replace an arbitrary $c \in \mathcal{K}^{(n)}$ by an arbitrary $c^{\text{new}} \notin \mathcal{K}^{(n)}$ it applies $\|\vec{y}^{(n)} - \vec{\mu}_{c^{\text{new}}}\| > \|\vec{y}^{(n)} - \vec{\mu}_c\|$ such that by the virtue of Proposition 1 $\mathcal{F}(\mathcal{K}, \Theta)$ decreases. As any arbitrary such replacement (any change of \mathcal{K}), results in a decrease of the free-energy, $\mathcal{F}(\mathcal{K}, \Theta)$ is maximized if \mathcal{K} contains the C' closest clusters.

□

Based on Corollary 1, we can now formulate a TV-EM algorithm for GMM (6) and (7) as is given in Alg. 3. For reasons that we will discuss further below, Alg. 3 will be referred to as k -means- C' .

3 k -means and TV-EM for GMMs

TV-EM for GMMs (Alg. 3) increases the similarity between k -means and standard EM for GMMs in two ways: (A) it relates Euclidean distances to a variational free-energy and thus to the GMM likelihood; and (B) it introduces ‘hard’ zeros in the updates of model parameters (some or many $q_c^{(n)}$ are zero). Crucial remaining differences are, however, (A) the weighted updates of the cluster centers in (10) compared to the k -means update and (B) the

Algorithm 3: The k -means- C' algorithm.

set $|\mathcal{K}^{(n)}| = C'$ for all n and init $\vec{\mu}_{1:C}$ and σ^2 ;

repeat

for $n = 1, \dots, N$ **do**

 define $\mathcal{K}^{(n)}$ such that

$\forall c \in \mathcal{K}^{(n)} \quad \forall \tilde{c} \notin \mathcal{K}^{(n)}:$

$\|\vec{y}^{(n)} - \vec{\mu}_c\| < \|\vec{y}^{(n)} - \vec{\mu}_{\tilde{c}}\|;$

 compute $q_c^{(n)}$ for all c and n using (9);

 update $\vec{\mu}_{1:C}$ and σ^2 with (10) and (11);

until $\vec{\mu}_{1:C}$ and σ^2 have converged;

update of the cluster variance σ^2 (11) along with the cluster centers for Alg. 3 which does not have a correspondence in k -means. By considering the first difference, the obvious next step is to consider a boundary case of Alg. 3 by demanding that the sets $\mathcal{K}^{(n)}$ shall contain just one element, i.e., we set $C' = 1$. Note in this respect that all proofs and considerations above apply for all $1 \leq C' \leq C$. While we recover for $C' = C$ standard EM for the GMM (6) and (7), we find that for $C' = 1$ the standard k -means algorithm is recovered.

Proposition 2

Consider the TV-EM algorithm (Alg. 3) for the GMM (6) and (7) with $\sigma^2 > 0$. If we set $C' = 1$, then the TV-EM updates of the cluster centers $\vec{\mu}_c$ (10) become independent of the variance σ^2 and are given by the standard k -means algorithm in Alg. 1.

Proof

If we choose $|\mathcal{K}^{(n)}| = C' = 1$ for all n , then each $\mathcal{K}^{(n)}$ computed in the TV-E-step of Alg. 3 contains according to Corollary 1 the index of the cluster center closest to $\vec{y}^{(n)}$ as only element. If we denote these centers by $c_o^{(n)}$, we get $\mathcal{K}^{(n)} = \{c_o^{(n)}\}$ and obtain for the truncated responsibilities $q_c^{(n)}$ in (9):

$$\begin{aligned} q_c^{(n)} &= \frac{\exp\left(-\frac{1}{2\sigma^2}\|\vec{y} - \vec{\mu}_c\|^2\right)}{\sum_{c' \in \{c_o^{(n)}\}} \exp\left(-\frac{1}{2\sigma^2}\|\vec{y} - \vec{\mu}_{c'}\|^2\right)} \delta(c = c_o^{(n)}) \\ &= \begin{cases} 1 & \text{if } c = c_o^{(n)} \\ 0 & \text{otherwise} \end{cases}, \end{aligned} \quad (14)$$

which is identical to $s_c^{(n)}$ in Alg. 1. By using $q_c^{(n)} =$

$s_c^{(n)}$ for the M-step, we consequently obtain:

$$\vec{\mu}_c^{\text{new}} = \frac{\sum_{n=1}^N s_c^{(n)} \vec{y}_n}{\sum_{n=1}^N s_c^{(n)}} \quad (15)$$

$$\sigma_{\text{new}}^2 = \frac{1}{DN} \sum_{n=1}^N \sum_{c=1}^C s_c^{(n)} \|\vec{y}^{(n)} - \vec{\mu}_c^{\text{new}}\|^2 \quad (16)$$

Now observe that the computation of $q_c^{(n)} = s_c^{(n)}$ and the updates of the $\vec{\mu}_c$ do not involve the parameter σ^2 . The cluster centers $\vec{\mu}_c$ can thus be optimized without requiring knowledge about the cluster variances σ^2 , i.e., the $\vec{\mu}_c$ optimization becomes independent of σ^2 . As the TV-EM updates for $q_c^{(n)}$ and $\vec{\mu}_c$ are identical to the updates of $s_c^{(n)}$ and $\vec{\mu}_c$ in Alg. 1, the optimization procedure for the $\vec{\mu}_c$ is given by the standard k -means algorithm. \square

A direct consequence of Proposition 2 is that standard k -means provably monotonously increases the truncated free-energy (12) with $C' = 1$. Notably, only for this choice of C' the updates of cluster means and variances decouple. We can, of course, add the variance updates to standard k -means but this does not effect the $\vec{\mu}_c$ updates. With or without σ^2 updates the free-energy monotonously increases. If our goal is the maximization of the free-energy objective, the σ^2 updates should be included, however. According to the independence of $\vec{\mu}_c$ optimization from σ^2 , it would be sufficient to update σ^2 once and only after k -means has optimized the cluster centers.

Proposition 2 shows that k -means is obtained from a variational free-energy objective. This free-energy is in turn closely related to the likelihood objective of GMMs (5). By analyzing the free-energy for $C' = 1$ more closely, we can make this relation more explicit and finally link the free-energy objective to the standard k -means objective (1).

Proposition 3

Consider a set of N data points $\vec{y}^{(1:N)} \in \mathbb{R}^D$ and the k -means algorithm (Alg. 1) where $s_{1:C}^{(1:N)}$ and $\vec{\mu}_{1:C}$ denote, respectively, the cluster assignments and cluster centers computed in one iteration. Further, let σ^2 denote the variance computed with $s_{1:C}^{(1:N)}$ and $\vec{\mu}_{1:C}$ as in (11),

$$\begin{aligned} \sigma^2 &= \sigma^2(s_{1:C}^{(1:N)}, \vec{\mu}_{1:C}) \\ &= \frac{1}{DN} \sum_{n=1}^N \sum_{c=1}^C s_c^{(n)} \|\vec{y}^{(n)} - \vec{\mu}_c\|^2. \end{aligned} \quad (17)$$

It then follows that each k -means iteration monotonously increases the free-energy $\mathcal{F}(s_{1:C}^{(1:N)}, \vec{\mu}_{1:C})$ given by:

$$\mathcal{F}(s_{1:C}^{(1:N)}, \vec{\mu}_{1:C}) = -\log(C) - \frac{D}{2} \log(2\pi e \sigma^2), \quad (18)$$

where e is Euler's number. The free-energy (18) is a lower bound of the GMM log-likelihood given by

$$\begin{aligned} \mathcal{L}(s_{1:C}^{(1:N)}, \vec{\mu}_{1:C}) &= \frac{1}{N} \sum_{n=1}^N \log \left(\frac{1}{C} \sum_{c=1}^C (2\pi \sigma^2)^{-\frac{D}{2}} \right. \\ &\quad \times \exp \left(-\frac{1}{2\sigma^2} \|\vec{y}^{(n)} - \vec{\mu}_c\|^2 \right) \Big). \end{aligned} \quad (19)$$

The difference between log-likelihood (19) and free-energy (18), $\mathcal{L} - \mathcal{F}$, is given by:

$$\begin{aligned} D(s_{1:C}^{(1:N)}, \vec{\mu}_{1:C}) &= \frac{D}{2} + \frac{1}{N} \sum_{n=1}^N \log \left(\sum_{c=1}^C \exp \left(-\frac{\|\vec{y}^{(n)} - \vec{\mu}_c\|^2}{2\sigma^2} \right) \right). \end{aligned} \quad (20)$$

If for all data points n applies $\frac{1}{\sigma^2} \|\vec{y}^{(n)} - \vec{\mu}_c\|^2 \ll \frac{1}{\sigma^2} \|\vec{y}^{(n)} - \vec{\mu}_{\tilde{c}}\|^2$ for the assigned cluster c compared to all non-assigned cluster \tilde{c} , then the free-energy bound becomes tight.

Proof

In the case $|\mathcal{K}^{(n)}| = C' = 1$ (which recovers k -means from Alg. 3) each $\mathcal{K}^{(n)}$ only contains one cluster which is given by the cluster assignments $s_c^{(n)}$ as: $\mathcal{K}^{(n)} = \{c | s_c^{(n)} = 1\}$. If we abbreviate this cluster for n with $c_o^{(n)}$, then it follows for the free-energy (12) after one k -means iteration:

$$\begin{aligned} \mathcal{F}(\mathcal{K}, \Theta) &= \frac{1}{N} \sum_n \log \left(\sum_{c \in \{c_o^{(n)}\}} \frac{1}{C} \mathcal{N}(\vec{y}^{(n)}; \vec{\mu}_c, \sigma^2 \mathbb{I}) \right) \\ &= \frac{1}{N} \sum_n \log \left(\frac{1}{C} \mathcal{N}(\vec{y}^{(n)}; \vec{\mu}_{c_o^{(n)}}, \sigma^2 \mathbb{I}) \right) \\ &= -\log(C) - \frac{D}{2} \log(2\pi \sigma^2) \\ &\quad - \frac{1}{2\sigma^2} \frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C s_c^{(n)} \|\vec{y}^{(n)} - \vec{\mu}_c\|^2, \end{aligned} \quad (21)$$

where we inserted the Gaussian density and then used $f(c_o^{(n)}) = \sum_c s_c^{(n)} f(c)$. σ^2 and $\vec{\mu}_c$ are the parameters obtained after a single k -means iteration. Following (11) we can therefore insert the expression $\frac{1}{DN} \sum_{n=1}^N s_c^{(n)} \|\vec{y}^{(n)} - \vec{\mu}_c\|^2$ for σ^2 , noting that the $\vec{\mu}_c$ are the same as in (21). The last term of (21) then simplifies to $-\frac{D}{2}$. If we now rewrite this as $-\frac{D}{2} \log(e)$ and combine with the second summand, we obtain

(18). The log-likelihood (19) is given directly by using the GMM (6) and (7) for the likelihood.

As k -means is a truncated variational approximation with $|\mathcal{K}^{(n)}| = C' = 1$ (Prop. 2), it follows (A) that the free-energy (18) is a lower bound of (19), and (B) that the bound monotonously increases in each k -means iteration. The former can only be seen considering Prop. 2. The latter can also directly be inferred by observing that σ^2 in (17) is the k -means objective of Alg. 1. As k -means monotonously decreased the objective, $-\log(\sigma^2)$ monotonously increases.

The difference between log-likelihood and free-energy can be derived from the KL-divergence $D_{\text{KL}}(q_c^{(n)}, r_c^{(n)})$. Using results of (Lücke, 2016) the KL-divergence for a truncated distribution is given by: $D_{\text{KL}}(q_c^{(n)}, r_c^{(n)}) = -\sum_n \log(\sum_{c \in \mathcal{K}^{(n)}} r_c^{(n)})$. Inserting $r_c^{(n)}$ (Alg. 2) for the GMM (6) and (7) we obtain:

$$\begin{aligned} D_{\text{KL}}(q_c^{(n)}, r_c^{(n)}) &= -\frac{1}{N} \sum_n \log \left(\sum_{c \in \mathcal{K}^{(n)}} \frac{\exp \left(-\frac{1}{2\sigma^2} \|\vec{y} - \vec{\mu}_c\|^2 \right)}{\sum_{c'} \exp \left(-\frac{1}{2\sigma^2} \|\vec{y} - \vec{\mu}_{c'}\|^2 \right)} \right) \\ &= -\frac{1}{N} \sum_n \log \left(\exp \left(-\frac{1}{2\sigma^2} \|\vec{y} - \vec{\mu}_{c_o^{(n)}}\|^2 \right) \right) \\ &\quad + \frac{1}{N} \sum_n \log \left(\sum_{c'} \exp \left(-\frac{1}{2\sigma^2} \|\vec{y} - \vec{\mu}_{c'}\|^2 \right) \right) \\ &= \frac{1}{2N\sigma^2} \sum_n \sum_c s_c^{(n)} \|\vec{y} - \vec{\mu}_c\|^2 \\ &\quad + \frac{1}{N} \sum_n \log \left(\sum_{c'} \exp \left(-\frac{1}{2\sigma^2} \|\vec{y} - \vec{\mu}_{c'}\|^2 \right) \right) \\ &= \frac{D}{2} + \frac{1}{N} \sum_n \log \left(\sum_{c'} \exp \left(-\frac{1}{2} \frac{\|\vec{y} - \vec{\mu}_{c'}\|^2}{\sigma^2} \right) \right), \end{aligned}$$

where we have again inserted the expression for σ^2 . The last line of the derivation above is the difference $D(s_{1:C}^{(1:N)}, \vec{\mu}_{1:C})$ given in (20). Finally, considering this expression, we observe that $D(s_{1:C}^{(1:N)}, \vec{\mu}_{1:C})$ goes to zero, if for each n the summand with smallest $\frac{1}{\sigma^2} \|\vec{y}^{(n)} - \vec{\mu}_c\|^2$ dominates the summation over c (as exp and log cancel in that case). \square

Considering Prop. 3, we can condense the relation between the likelihood objective $\mathcal{L}(\Theta)$ of GMM (6) and (7) and the standard k -means objective (1) as follows:

$$\mathcal{L}(\Theta) \geq -\log(C) - \frac{D}{2} \log \left(\frac{2\pi e}{DN} \mathcal{J}(\vec{\mu}_{1:C}) \right), \quad (22)$$

where we replaced σ^2 in (17) by (1). Eqn. 22 is, however, a slightly imprecise formulation as we here implicitly assume that $\Theta = (\vec{\mu}_{1:C}, \sigma^2)$ is updated ac-

cording to k -means together with variance update (16). Also note that the relation $\mathcal{L} \geq \mathcal{F}$ has only been derived for the state of $s_{1:c}^{(1:N)}$ and $\tilde{\mu}_{1:C}$ after any one iteration. Prop. 3, which also includes (20), is hence more detailed and accurate. Eqn. 22 (and similar formulations of Eqns. 19 and 20, see Appendix) may be appealing as it highlights the direct link between two very well-known clustering objectives, and as it is directly applicable in practice.

4 Application of Theoretical Results

We can now apply our results to derive possible generalizations and to study their relations to previously suggested k -means generalizations. When considering Alg. 3 the most straight-forward generalizations of k -means are given (A) by using $C' > 1$ which can allow for more than one winner per data point; and (B) by changing the criterion (13) for cluster selection in Prop. 1.

4.1 Generalization k -means- C'

Using any $C' > 1$, we know because of Corollary 1 that Alg. 3 increases a free-energy bound of the log-likelihood w.r.t. GMM (6) and (7), and we furthermore know because of (12) that increasing the size of \mathcal{K} increases $\mathcal{F}(\mathcal{K}, \Theta)$. The free-energy with $C' > 1$ is thus a tighter likelihood bound than the k -means free-energy (18). On the other hand, and in contrast to k -means, the cluster centers $\tilde{\mu}_{1:C}$ can not be computed in isolation anymore, as the truncated posteriors $q_c^{(n)}$ in Eqn. 9 require for $C' > 1$ the variances σ^2 . Furthermore, the free-energy $\mathcal{F}(\mathcal{K}, \Theta)$ does not take on the compact form (18). The additional computational overhead for $C' > 1$ is however relatively limited compared to standard k -means (however, it may be more difficult to derive accelerated versions; Arthur and Vassilvitskii, 2006; Bachem et al., 2016). Furthermore, we can simplify the expression for the free-energy (12) using entropy limits (Lücke and Henniges, 2012).

Proposition 4

Consider a set of N data points $\vec{y}^{(1:N)} \in \mathbb{R}^D$ and let the cluster centers $\tilde{\mu}_{1:C}$ and the variance σ^2 be updated according to Alg. 3 with $1 \leq C' \leq C$. Furthermore, let $q_{1:C}^{(1:N)}$ be the truncated responsibilities

(9) computed with C' . If we take the $q_{1:C}^{(1:N)}$ and the values $\tilde{\mu}_{1:C}$ and σ^2 to be those computed in one EM iteration, then the lower free-energy bound of the log-likelihood (19) after the iteration is given by:

$$\mathcal{F}(q_{1:C}^{(1:N)}, \tilde{\mu}_{1:C}) = -\log(C) - \frac{D}{2} \log(2\pi e \sigma^2) - \frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C q_c^{(n)} \log(q_c^{(n)}), \quad (23)$$

with σ^2 as in (17) but $s_c^{(n)}$ replaced by $q_c^{(n)}$.

Proof

Given GMM (6) and (7), we observe that the entropy $\mathcal{H}(p(\vec{y}|c, \Theta)) = \mathcal{H}(\mathcal{N}(\vec{y}; \tilde{\mu}_c, \sigma^2 \mathbb{1}))$ of the noise distribution does not change with c . The GMM therefore has an entropy limit (Lücke and Henniges, 2012) given by:

$$\begin{aligned} \overline{\mathcal{Q}}(\Theta) &= -\mathcal{H}(p(c|\Theta)) - \mathcal{H}(p(\vec{y}|c, \Theta)) \\ &= -\log(C) - \frac{D}{2} \log(2\pi e \sigma^2), \end{aligned}$$

which is derived simply by inserting (6) and (7). If we reformulate (following Lücke and Henniges, 2012) the free-energy (12) such that it is expressed in terms of this entropy limit, we obtain:

$\mathcal{F}(\mathcal{K}, \Theta) = \overline{\mathcal{Q}}(\Theta) + \frac{D}{2} \left(1 - \frac{\sigma_{\text{new}}^2}{\sigma^2}\right) + \frac{1}{N} \sum_n \mathcal{H}(q_c^{(n)})$ where σ_{new}^2 is given by (11). If $\mathcal{F}(\mathcal{K}, \Theta)$ is computed after an EM iteration of Alg. 3, then σ^2 is equal to σ_{new}^2 and we obtain (23), which can be computed in terms of $q_{1:C}^{(1:N)}$ and $\tilde{\mu}_{1:C}$. \square

By considering Prop. 4, we observe that (23) differs from the k -means free-energy (18) by one additional entropy term. In other words, the larger value of $\mathcal{F}(q_{1:C}^{(1:N)}, \tilde{\mu}_{1:C})$ for $C' > 1$ compared to k -means ($C' = 1$) is given by the average entropy of the non-binary truncated distributions $q_c^{(n)}$ of k -means- C' . If we set $C' = 1$ for (23), we recover (18) because the entropy of binary truncated posteriors is zero.

4.2 Other State Selection Criteria

Other than changing C' to other values, Prop. 1 sets the ground to also change the selection criterion for states in \mathcal{K} . Let us first consider *lazy k-means* which is a relatively recent k -means generalization used to study convergence properties (Har-Peled and Sadri,

2005). Lazy k -means only reassigns a data point n from a cluster \tilde{c} to a new cluster c if:

$$(1 + \epsilon) \|\vec{y}^{(n)} - \vec{\mu}_c\| < \|\vec{y}^{(n)} - \vec{\mu}_{\tilde{c}}\|, \quad (24)$$

where ϵ is a small non-negative constant. For $\epsilon = 0$ standard k -means is recovered.

By considering Prop. 1, any replacement of states in \mathcal{K} according to (24) would also increase the free-energy (12). Based on our variational interpretation, lazy k -means corresponds to a partial TV-E-step. In analogy to Prop. 1 we can show that (12) is monotonously increased but it is not necessarily maximized – Corollary 1 does not apply. However, the essential observation of a decoupled $\vec{\mu}$ and σ^2 update only depends on C' being equal to one. Prop. 2 thus generalizes to the lazy k -means case and can now be stated as follows:

Proposition 5

Consider the TV-EM algorithm (Alg. 3) but with (24) instead of the pure Euclidean distance. If we set $C' = 1$, then the TV-EM updates of the cluster centers $\vec{\mu}_c$ (10) become independent of the variance σ^2 and are given by the lazy k -means algorithm..

Proof

The proof is analogous to the one of Prop. 2 only that $c_o^{(n)}$ is now the cluster of $\vec{y}^{(n)}$ for which applies: $\forall \tilde{c} \neq c_o^{(n)} : \|\vec{y}^{(n)} - \vec{\mu}_{\tilde{c}}\| < (1 + \epsilon) \|\vec{y}^{(n)} - \vec{\mu}_c\|$. The cluster assignments thus become those of lazy k -means, while the forms of the parameter updates remain those of standard k -means (i.e., the same as used for lazy k -means).

□

For lazy k -means, note that Prop. 3 can be generalized to lazy k -means in the same way as Prop. 5 generalizes Prop. 2. Basically, the binary cluster assignments $s_c^{(n)}$ change according to (24). This means that Eqns. 18 to 20 and Eqn. 22 can be used if the k -means cluster assignments are everywhere replaced by the lazy k -means cluster assignments.

Lazy k -means was introduced because polynomial complexity bounds for its running time can be derived (Har-Peled and Sadri, 2005). Already by considering standard k -means we can conclude from Prop. 3 that the lower bound (18) strictly increases because k -means strictly increases objective (1). Furthermore, using the exponential running time bounds of k -means (e.g. Inaba and Katoh, 2000), we can con-

clude that the bound is maximized in exponential but finite time. Our variational considerations without the derived relation to the k -means objective, do not allow us to conclude a strict increase and a finite running time. Instead of an exponential running time, lazy k -means has a low polynomial bound which depends on ϵ (Har-Peled and Sadri, 2005), and we can conclude that the bound (22) with \mathcal{J} computed using the lazy k -means assignments is increased in polynomial time.

Finally, note that the criterion in Alg. 2 can also be changed according to a generalization of the underlying generative model itself. If we use the GMM (2) and (3), for instance, we can show that the free-energy that corresponds to the log-likelihood (5) is increased if we replace a cluster $\tilde{c} \in \mathcal{K}^{(n)}$ with a cluster $c \notin \mathcal{K}^{(n)}$ such that:

$$\begin{aligned} & \|\vec{y} - \vec{\mu}_c\|_{\Sigma_c}^2 + \log(|2\pi\Sigma_c|) - 2\log(\pi_c) \\ & < \|\vec{y} - \vec{\mu}_{\tilde{c}}\|_{\Sigma_{\tilde{c}}}^2 + \log(|2\pi\Sigma_{\tilde{c}}|) - 2\log(\pi_{\tilde{c}}). \end{aligned} \quad (25)$$

Again, we can derive an algorithm for $C' = 1$. In this case, we can in contrast to the criterion (24) generalize Corollary 1, i.e., criterion (25) corresponds to a full TV-E-step which maximizes the free-energy (for fixed Θ). On the other hand, neither Prop. 2 or 3 have straight-forward generalizations for (25) because of the change of the underlying generative model.

4.3 Other k -means generalizations

Considering the above discussed generalizations, note that the lazy k -means generalization as well as the generalization based on (25) can be combined with generalizations for $C' > 1$. All these generalizations may be interesting to explore both because of their theoretical properties and because of their potential practical usefulness. In the set of all these generalizations ($\epsilon \geq 0$, $C' \geq 1$, different GMMs) note that k -means is special: Only for $\epsilon = 0$ in (24) the free-energy is fully maximized, and of all the considered generalizations for $C' > 1$ or generalized GMMs (25) only for k -means the $\vec{\mu}_c$ optimization becomes independent of cluster variance parameters.

Because of its popularity, there are many generalizations of k -means other than those discussed here so far. Well known examples other than lazy k -means are, for instance, soft- k -means (MacKay, 2003) or fuzzy- k -means (e.g. Bezdek, 1981). Soft- k -means was suggested as a ‘non-hard’ k -means generalization and

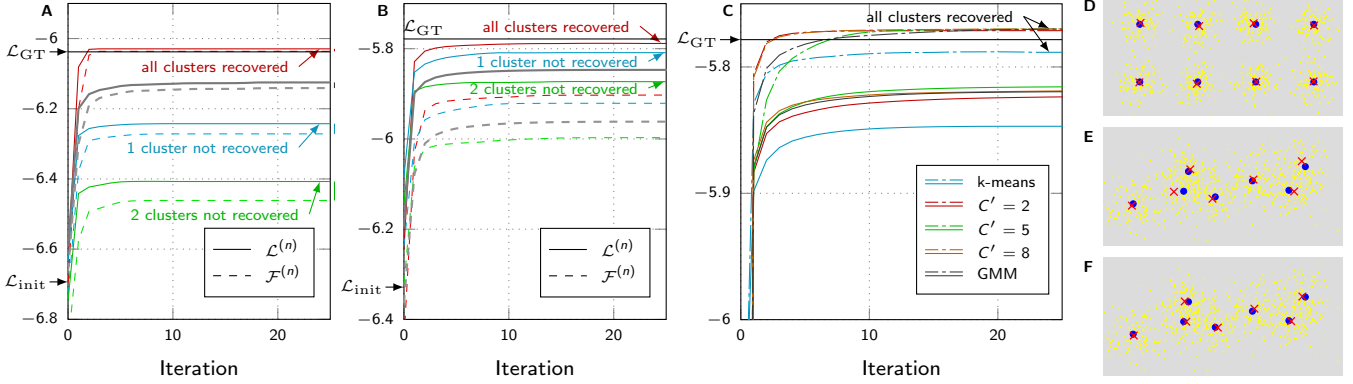


Figure 1:

A and B show the likelihood and free energy per iteration for three individual runs (red, blue and green) and the mean of 100 independent runs (gray) of Alg. 1 on a BIRCH data set with grid-like clusters (A and D) and uniform randomly chosen clusters (B, C, E and F). C shows the mean loglikelihood (solid line) and of the best run (dashed) of 100 runs of Alg. 3 for different C' . Visualization of some ground truth cluster centers (blue dots) and found cluster centers of the best runs (red crosses) are shown in D, E (for k -means) and F (for k -means- C' with $C' = 2$).

its stiffness parameter β closely links to the cluster variances (essentially $\beta = \frac{1}{2\sigma^2}$) of GMM (6) and (7). The variational generalizations discussed here for $C' > 1$ make k -means ‘softer’ by allowing more than one non-zero value for the cluster assignments. This is different from soft k -means which maintains non-zero values for all cluster assignments. However, problems with sensitivity to stiffness values and sensitivity to initial conditions compared to standard k -means (Barbakh and Fyfe, 2008) may be related to Prop. 1 and Prop. 2 which imply that for soft assignments σ^2 updates should not be neglected. Soft k -means could hence be improved by updating the stiffness, which would further increase its similarity to GMM clustering.

An alternative to soft k -means are *fuzzy k -means* approaches, suggested earlier (e.g. Bezdek, 1981; Yang, 1993, for references). Fuzzy k -means can be described as generalization of the k -means objective (1) by using non-binary $s_c^{(n)}$ in the place of the k -means assignments. Fuzzy k -means algorithms then update weighted cluster assignments and cluster centers in order to minimize such objectives. Fuzzy k -means is different from truncated variational generalizations as can be seen by considering the log-likelihood and the free-energy objectives derived here. Most notably, Prop. 4 shows that the average entropy of the cluster assignments emerges as term in addition to a softened objective in the context of GMM likelihood optimization. Fuzzy k -means algorithms optimize objectives without assignment entropy, and are

therefore different from any of the algorithms considered here. Soft k -means, on the other hand, is more closely related than fuzzy k -means as it can be shown to be derivable from an objective including an entropy term (e.g. Kim et al., 2007).

5 Numerical Verification

Before we conclude, we briefly numerically verify the main theoretical results of this work. We use a BIRCH dataset with $C = 25$ clusters on a 5×5 grid with $N = 100$ data points per cluster (same data set for all runs) as partly shown in Fig. 1D. Fig. 1A shows different runs of standard k -means and the time course of the free-energy and likelihood computed using (18) and (19) of Prop. 3, respectively. The shown exemplary runs converge to different optima. The run with highest final free-energy recovers all cluster centers and results in a log-likelihood larger than the log-likelihood of the generating (ground-truth) parameters. We verified that this (small) overfitting effect decreases with increasing N . The bound for the best run is relatively tight, which is consistent with (20) of Prop. 3 for small σ^2 . The gap is larger for local optima, which have to have a larger σ^2 according to (18) and consequently higher entropy for $q_c^{(n)}$ of $C' > 1$ including $C' = C$. The same applies for clusters with increased overlap in Fig. 1C,E,F, where we use the same setting as for Fig. 1A but with uniformly distributed generating cluster centers as seen in Fig. 1E and 1F. Note that we use the seeding of k -

means++ (Arthur and Vassilvitskii, 2006) for Fig. 1. The initial values are thus already relatively good (see $\mathcal{L}_{\text{init}}$).

Fig. 1C shows different runs of k -means- C' for the data set of Fig. 1B. Using k -means- C' with different numbers of winning clusters C' can prevent shifted cluster centers caused by unsymmetrical cluster overlaps (compare Fig. 1E and 1F). Final likelihoods of the best runs with $C' > 1$ can hence be higher (consistent with Prop. 4) but the best k -means ($C' = 1$) likelihoods are not much lower and initially increase faster.

6 Conclusion

The here studied quantitative link between the k -means objective and the GMM log-likelihood allows for relating many of the studies on different algorithmic properties of k -means and GMM results. One example application that has been discussed here in more detail was the analysis of time complexity of k -means and lazy k -means. Notably, any time complexity bound for k -means and lazy k -means can be translated into a bound on the log-likelihood, e.g., by using Props. 3 and 5. Similar relations can presumably be derived for a smoothed analysis of k -means (e.g. Arthur et al., 2009) in future applications of the results derived in this work. Likelihood bounds are, on the other hand, of interest for theoretical studies of GMM optimization (e.g. Kalai et al., 2010; Moitra and Valiant, 2010; Xu et al., 2016). Our results may thus serve as a valuable tool in combining complexity studies of k -means and GMM-based clustering. Of the many studies which consider k -means and data samples of GMMs (e.g. Chaudhuri et al., 2009, and cites therein), there is none, to our knowledge, that has derived k -means as variational approximation. Work by Pollard et al. (1981) is maybe one of the most relevant studies, as it proves a theorem which relates the convergence points of k -means to an underlying distribution. In the sense of a central limit theorem, this distribution is given by a GMM with clusters of specific covariance. Cluster overlap in the samples influences the cluster shapes via non-zero off-diagonal elements. The question of Pollard et al. (1981) is thus how to fit a GMM (in a central limit theorem sense) to correspond to k -means convergence points. Prop. 3 may be related to the theorem of Pollard et al. (1981) but a closer inspection would require a more extensive analysis.

As for clustering in general, k -means also remained of interest in the probabilistic Machine Learning community, and notably in the field of non-parametric approaches. Welling and Kurihara (2006) suggested “Bayesian k -means”, for instance, and used variational Bayesian approximations in order to obtain k -means-like running time behavior for model selection. Later on, Kulis and Jordan (2011) followed a Bayesian treatment, and combined it with the relation of k -means to GMMs obtained in the limit of zero variances (see Sec. 1). In this way they derived new ‘hard assignment’ algorithms based on a Gibbs sampler used within a non-parametric approach. Also more generally, the limit to zero cluster variance remained the most well known relation between k -means and GMMs. From the probabilistic point of view, this limit is unsatisfactory, however, as the likelihood of data points under a GMM with ($\sigma^2 \rightarrow 0$) also approaches zero. By applying truncated variational distributions we in this study can maintain any finite variance $\sigma^2 > 0$. Truncated approaches, which allowed finite variance $\sigma^2 > 0$, are novel compared to standard variational approaches which assume a-posteriori independence (e.g. Saul et al., 1996; Jaakkola, 2000). Truncated EM approaches (Lücke and Eggert, 2010; Sheikh et al., 2014; Lücke, 2016) aim at scalable and accurate approximations without assuming a-posteriori independence, a goal it shares with many later approaches (e.g. Rezende and Mohamed, 2015; Salimans et al., 2015; Kucukelbir et al., 2016). Truncated EM is the natural variational approximation for k -means-like algorithms, and is here not only related but becomes, indeed, identical to standard k -means.

One of the implications following from this observations is a direct quantitative link between the k -means and GMM objectives (e.g., Eqn. 22), which we believe may be of potentially high relevance for further theoretical and empirical studies.

Appendix

Appendix A

Using (1) instead of σ^2 we obtain for Eqns. 18 to 20:

$$\mathcal{F}(\vec{\mu}_{1:C}) = -\log(C) - \frac{D}{2} \log \left(\frac{2\pi e}{DN} \mathcal{J}(\vec{\mu}_{1:C}) \right), \quad (26)$$

$$\mathcal{L}(\vec{\mu}_{1:C}) = \mathcal{F}(\vec{\mu}_{1:C}) + D(\vec{\mu}_{1:C}), \quad (27)$$

$$D(\vec{\mu}_{1:C}) = \frac{D}{2} + \frac{1}{N} \sum_{n=1}^N \log \left(\sum_{c=1}^C \exp \left(-\frac{D}{2} \frac{N \|\vec{y}^{(n)} - \vec{\mu}_c\|^2}{\mathcal{J}(\vec{\mu}_{1:C})} \right) \right), \quad (28)$$

where the k -means objective $\mathcal{J}(\vec{\mu}_{1:C})$ is using the values for $s_{1:C}^{(1:N)}$ and $\vec{\mu}_{1:C}$ as they have been computed in one k -means iteration (Alg. 1), i.e., $s_{1:C}^{(1:N)}$ has been computed with the cluster centers of the preceding iteration, while $\vec{\mu}_{1:C}$ are the current values. In that case $\mathcal{J}(\vec{\mu}_{1:C}) = DN\sigma^2$ with σ^2 given by (17). The log-likelihood (19) we can based on Prop. 3 express as the sum of the two other entities. In summary we obtain Eqns. 26 to 28.

The inequality (22) is obtained as $D(\vec{\mu}_{1:C})$ is larger than zero. Formula (22) can also be interpreted a bit more generally if we consider that a general GMM (2) to (3) contains the GMM with same mixing coefficients and isotropic Gaussians (6) to (7) as special case. If we interpret $\mathcal{L}(\Theta)$ as maximal value achievable with a general GMM (excluding generate cases converging to infinite values), than any $\mathcal{F}(\vec{\mu}_{1:C})$ value computed with k -means represents a lower-bound of such maximally achievable likelihoods. As mixture models can be very efficient in matching data distributions, any k -means would also set a likelihood bound for any other likelihood based probabilistic approach.

Note that once the cluster means have converged, the rather technical point that $\mathcal{J}(\vec{\mu}_{1:C})$ has to use the $s_c^{(n)}$ and $\vec{\mu}_c$ values as computed in a previous k -means iteration does vanish. The formulas (26) to (28) can thus after convergence be applied without considering the order of the computation of $s_{1:C}^{(1:N)}$ and $\vec{\mu}_{1:C}$. In practice an evaluation of a k -means run would hence use formulas (26) to (28) to compute likelihood and free-energy values for comparison.

Appendix B: k -means- Σ - π for general GMMs

Using criterion (25) the a k -means- Σ - π algorithm is given by Alg. 4. As we proceeded for k -means- C' , the

Algorithm 4: k -means- Σ - π .

init Θ of (2) to (3);

repeat

for $n = 1, \dots, N$ and $c = 1, \dots, C$ **do**

 compute assignments $t_c^{(n)}$ using (25);

for $c = 1, \dots, C$ **do**

 update Θ as in (GMM 1) to (GMM 3)

 using $t_c^{(n)}$ in place of $r_c^{(n)}$;

until Θ have converged;

replacement of $r_c^{(n)}$ by $t_c^{(n)}$ is possible by the virtue of (Lücke, 2016). Alg.4 is a fusion of Alg. 1 and Alg. 2.

References

- Arthur, D., Manthey, B., and Röglin, H. (2009). k-means has polynomial smoothed complexity. In *IEEE Symp. Foundations of Comp. Sci.*, pages 405–414.
- Arthur, D. and Vassilvitskii, S. (2006). How slow is the k-means method? In *Comp. Geo.*, pages 144–153.
- Bachem, O., Lucic, M., Hassani, H., and Krause, A. (2016). Fast and provably good seedings for k-means. In *NIPS*, pages 55–63.
- Barbakh, W. and Fyfe, C. (2008). Online clustering algorithms. *International Journal of Neural Systems*, 18(03):185–194.
- Belkin, M. and Sinha, K. (2010). Polynomial learning of distribution families. In *Symp. Comp. Sci.*, pages 103–112.
- Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer.
- Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Springer.
- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Chaudhuri, K., Dasgupta, S., and Vattani, A. (2009). Learning mixtures of gaussians using the k-means algorithm. *arXiv preprint arXiv:0912.0086*.
- Dai, Z. and Lücke, J. (2014). Autonomous document cleaning. *TPAMI*, 36(10):1950–1962.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39:1–38.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley-Interscience (2nd Edition).
- Forster, D. and Lücke, J. (2017). Truncated variational EM for semi-supervised neural simpletrons. In *IJCNN 2017, accepted*. <http://arxiv.org/abs/1702.01997>.
- Har-Peled, S. and Sadri, B. (2005). How fast is the k-means method? *Algorithmica*, 41(3):185–202.
- Hughes, M. C. and Sudderth, E. B. (2016). Fast learning of clusters and topics via sparse posteriors. *arXiv preprint arXiv:1609.07521*.
- Inaba, M. and Katoh, N. (2000). Variance-based k-clustering algorithms by voronoi diagrams and randomization. *IEICE Trans. Inf. Sys.*, 83(6):1199–1206.
- Jaakkola, T. (2000). Tutorial on variational approximation methods. In Oppor, M. and Saad, D., editors, *Advanced mean field methods*. MIT Press.
- Kalai, A. T., Moitra, A., and Valiant, G. (2010). Efficiently learning mixtures of two gaussians. In *Proc. ACM Symp. Theo. Comp.*, pages 553–562. ACM.
- Kim, J., Shim, K.-H., and Choi, S. (2007). Soft geodesic kernel k-means. In *ICASSP*, volume 2, pages II–429.
- Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2016). Automatic differentiation variational inference. *CoRR*, abs/1603.00788.
- Kulis, B. and Jordan, M. I. (2011). Revisiting k-means: New algorithms via bayesian nonparametrics. *arXiv preprint arXiv:1111.0352*.
- Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Trans. Inf. Theory*, 28(2):129–137.
- Lücke, J. (2016). Truncated variational expectation maximization. *arXiv preprint, arXiv:1610.03113*.
- Lücke, J. and Eggert, J. (2010). Expectation truncation and the benefits of preselection in training generative models. *JMLR*, 11:2855–900.
- Lücke, J. and Henniges, M. (2012). Closed-form entropy limits. In *AISTATS*, volume 22, pages 731–740.
- Lücke, J. and Sahani, M. (2008). Maximal causes for non-linear component extraction. *JMLR*, 9:1227–67.
- MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge Univ. Press.

- McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, volume 84. Marcel Dekker.
- Moitra, A. and Valiant, G. (2010). Settling the polynomial learnability of mixtures of Gaussians. In *IEEE Symp. Found. Comp. Sci*, pages 93–102.
- Pollard, D. et al. (1981). Strong consistency of k -means clustering. *The Annals of Statistics*, 9(1):135–140.
- Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. *ICML*.
- Salimans, T., Kingma, D., and Welling, M. (2015). Markov chain monte carlo and variational inference: Bridging the gap. *ICML*.
- Saul, L. K., Jaakkola, T., and Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of artificial intelligence research*, 4(1):61–76.
- Sheikh, A.-S., Shelton, J. A., and Lücke, J. (2014). A truncated EM approach for spike-and-slab sparse coding. *JMLR*, 15:2653–2687.
- Shelton, J. A., Gasthaus, J., Dai, Z., Lücke, J., and Gretton, A. (2014). GP-select: Accelerating EM using adaptive subspace preselection. arxiv.org/abs/1412.3411 now in press for *Neural Computation*.
- Steinley, D. (2006). K-means clustering: A half-century synthesis. *Brit. J. Math. and Stat. Psych.*, 59(1):1–34.
- Welling, M. and Kurihara, K. (2006). Bayesian k-means as a maximization-expectation algorithm. In *Proc. SIAM Conf. Data Mining*, pages 474–478.
- Xu, J., Hsu, D. J., and Maleki, A. (2016). Global analysis of expectation maximization for mixtures of two gaussians. In *NIPS*, pages 2676–2684.
- Yang, M.-S. (1993). A survey of fuzzy clustering. *Mathematical and Computer modelling*, 18(11):1–16.